TRANSPARENCY INTERNATIONAL
the global coalition against corruption

# THE CORRUPTION RISKS OF ARTIFICIAL INTELLIGENCE

VISIT…

Transparency International is a global movement with one vision: a world in which government, business, civil society and the daily lives of people are free of corruption. With more than 100 chapters worldwide and an international secretariat in Berlin, we are leading the fight against corruption to turn this vision into reality.

**www.transparency.org**

Working paper

# The corruption risks of artificial intelligence

Author: Nils Christopher Köbis, Christopher Starke, Jaselle Edward-Gill

Reviewers: Matthew Jenkins, Jon Vrushi, Lars Wriedt, Daniel Eriksson



Funded by the European Union

This document should not be considered as representative of the European Commission's official position. Neither the European Commission, Transparency International nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information.

# INTRODUCTION

Machines powered with artificial intelligence (AI) technology are being increasingly deployed in our economic, political and social lives. AI is now widely seen as a transformative and disruptive technology for political and social order[1]. Three factors play a particularly important role in the growing influence of AI. First, computing machinery like microchips and graphics processing units has improved, enabling an increase in computing power. Second, more data is available in a digital format, providing the necessary fuel for AI systems. Finally, in the past decade, technological breakthroughs have occurred in the field of machine learning – a subdomain of AI that trains algorithms to engage in tasks autonomously.

## THE BRIGHT AND DARK SIDES OF ARTIFICIAL INTELLIGENCE

AI has been implemented in a wide range of domains, hinting at its positive potential to affect society. AI can contribute to faster, more efficient processes, invigorate the economy,[2] address environmental challenges,[3] and deliver breakthroughs in biological sciences, such as predicting the folding of proteins.[4] In the health sector, AI systems used for medical diagnostics and decision support have achieved and surpassed human expert levels across various tasks.[5] There has been rapid progress in language generation and understanding, a domain long considered an almost unconquerable bastion of human intelligence. Recently released large language models (LLMs) enable the production of human-like text ranging from poetry to news articles, as well as writing computer code or holding conversations via chatbots.[6]

The growing influence of AI in society has also led to the development of AI systems in the fight against corruption.[7] For instance, AI has been used to automatically predict corruption risks based on data from news media,[8] police archives,[9] and financial reports.[10] Moreover, tweetbots can publicise suspicious cases of reimbursement claims by parliamentarians and encourage their followers to investigate the cases further.[11] Algorithmic systems are already regularly used in the context of anti-money laundering, where they are employed to analyse massive datasets of financial transactions to spot irregularities. As such, they can flag specific transactions to be investigated further or even restrict transactions before they take place.[12]

However, AI can also have negative repercussions that are often seemingly unintentional. Introducing AI systems in both the private and public sector can produce undesirable outcomes as a result of biased input data, faulty algorithms or irresponsible implementation.[13] For instance, certain facial detection software has been shown to perform poorly on people of colour because it was not trained with sufficiently diverse training data sets.[14]

Besides such apparent unintended effects, ever more cases are being documented in which AI is intentionally weaponised.[15] To name a few examples, scammers have used AI based hyper-realistic imitations of audio-visual content called deepfakes for novel fraud schemes. In a single instance, they defrauded more than US$240,000 from a CEO who falsely believed to be speaking to the boss of the parent company.[16] AI algorithms can also be used as transformative tools for algorithmic collusion, such as where algorithms autonomously coordinate to fix prices in a collusive manner.[17] Moreover, scammers have used bots to spread disinformation about a small company to artificially boost the company's value, at which point they sold the practically worthless stocks.[18]

Yet, this malicious use of AI does not constitute corrupt AI because these acts are not executed by power holders. Power holders describe people with "access to resources that are needed or valued by others".[19] Indeed, the risk of entrusted power

holders abusing AI for private gains have been largely neglected.[20]

It is these risks that this paper addresses. It does so by:

i) outlining the key features of AI systems to map potential corruption risks;
ii) systematising key examples of how AI could be used in a corrupt fashion along the dimensions of design, manipulation and application of AI;
iii) highlighting how AI technology differs from classical digital technologies in enabling corruption; and iv) closing with some initial practical recommendations related to regulatory, technical and human factors.

These insights rest on two main pillars. First, we draw on an extensive review of the academic and grey literature, the latter including policy reports and media articles. Second, we conducted semi-structured interviews with experts in data science, corruption research, AI ethics and policymaking from the private and public sectors.

# CHARACTERISTICS OF ARTIFICAL INTELLGIENCE

As with defining corruption, providing an all-encompassing, widely agreed definition of AI is challenging. The European Union (EU) defines AI as "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals".[21]

Instead of engaging in the definitional debate, this report differentiates between key forms of AI. It emphasises the features that set AI technologies apart from classic information communication technologies (ICT) when used for corrupt purposes.

To understand the imminent risks of AI for corruption, we focus on the already existing forms of artificial narrow intelligence (compared to artificial general intelligence). While artificial general intelligence describes machines generally surpassing humans in various tasks, artificial narrow intelligence refers to machines that can perform a single narrowly defined task (or a small set of related tasks) at human levels. "Under the hood", such systems are powered by different types of algorithms.

## RULE-BASED VERSUS MACHINE LEARNING ALGORITHMS

The first class of algorithms that are often (although not always) classified as rule-based algorithms. As the name suggests, human programmers write codified rules that define all aspects of a specific task. An example of a rule-based algorithm is, "If event X occurs, then do Y; if an event other than X occurs, then do Z". It is important to note that the link between what the programmer specifies and what the algorithm does is entirely predictable. Rule-based algorithms can be corrupted if they are intentionally set up to favour a particular individual or group.

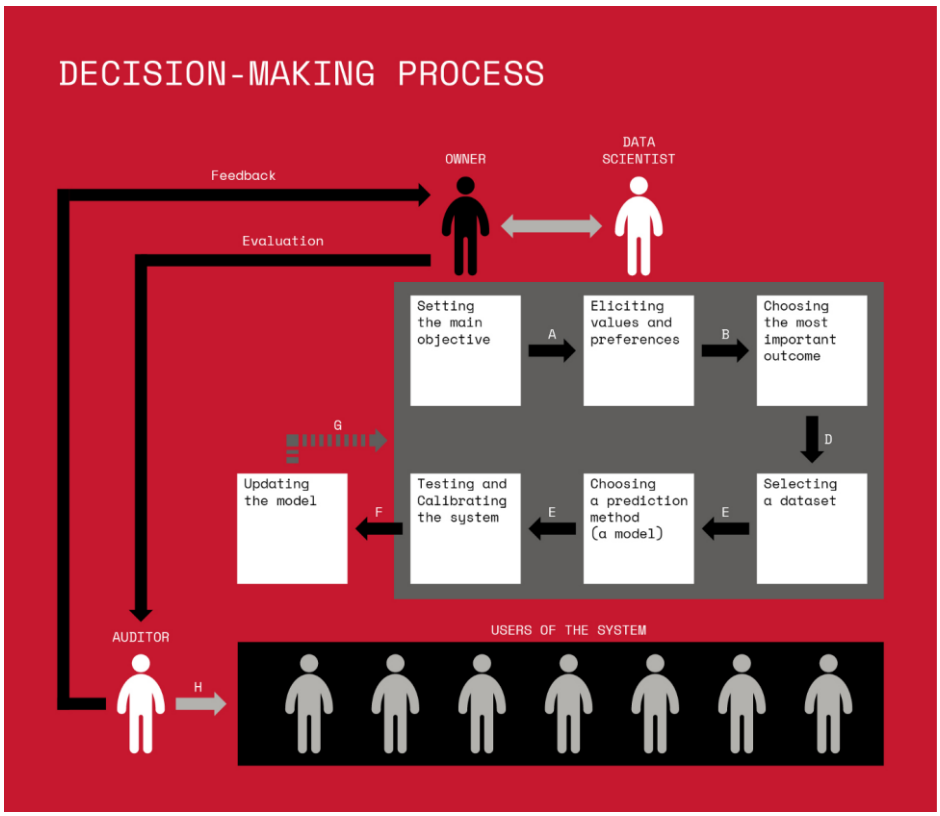
The second branch is machine learning algorithms. In machine learning, algorithms do not follow pre-specified rules but rather '"learn' from data or their own 'experiences"'.[22] A particularly innovative sub-branch of machine learning is deep learning**.** It uses so-called neural networks that loosely mimic the structure of the human brain. A neural network model usually entails many layers. Each layer, in turn, consists of simple processing units (called neurons) through which information is processed.[23] The interest and uptake of these advanced forms of AI in the private and public sectors is growing.[24] Due to their self-learning abilities, the detailed decision-making process of machine learning algorithms is often opaque. Even the programmers do not always fully understand how a particular decision was reached. The opacity of many machine learning systems makes them vulnerable to those who want to exploit them for corrupt purposes.

## THE IMPLEMENTATION PROCESS OF ARTIFICIAL INTELLIGENCE

Grasping the risks of corrupt AI requires a closer inspection of the socio-technical implementation of AI. In a simplified schematic depiction, the process of developing and implementing AI systems typically involves four actors: i) the commissioner or owner; ii) the data scientist; iii) the auditor; and iv) the user of the system (see Figure 1).

The commissioner, also known as the domain expert, initiates the process of developing a new instance of AI. The commissioner has expertise in the institutional context and sets the system's objectives. Think, for example, of a company's human resources department that plans to implement an AI system to autonomously pre-screen applicants. The commissioner could corrupt the AI systems by setting it up to illegitimately favour a particular group of applicants (in exchange for private gain, such as a bribe or kickback).

**Figure 1**. The decision-making process for designing and implementing AI systems, from Szymielewicz (2020).[25]




The commissioner then assigns data scientists and programmers with the technical expertise to design the desired algorithm. Ideally, this step entails iterative steps between the commissioner and the data scientist in defining the central goal of the system, agreeing on ethical values and preferences, identifying reliable and valid training data, and selecting a suitable prediction model.[26] These multiple feedback loops and continuous updating of the system are intended to ensure the system's security and fairness and mitigate potential unintended consequences. Yet, it is essential to note that this ideal clear-cut distinction between the commissioner and the data scientist does not always exist in reality. Both roles are often exerted by the same person, especially in small companies and start-ups. Similar to the corruption risk outlined above, also data scientists and programmers might embed biases into the AI system for personal gains.

A crucial step in implementing an AI system consists of rigorous audits by independent auditors. Auditors evaluate the prediction model for potential security concerns such as vulnerabilities for malicious abuse of the algorithm and feed the insights back to the commissioner and data scientist. Ideally, auditors also evaluate the broader institutional and societal context by, for example, identifying and mitigating issues such as unintended bias. However, in reality, audits primarily focus on the technical vulnerabilities of the AI system.[27] Therefore, the auditors require specialised technical expertise and are almost exclusively data scientists. While it is particularly important to ensure the ethical use of AI, in practice such code audits are infrequent and can potentially be corrupted if auditors turn a blind eye to potential harm caused by AI in exchange for bribes.

Finally, the implementation of AI systems involves the users directly affected by AI decision-making. These users can include laypeople who may not even be aware that they are subject to AI decision-making, such as job seekers whose applications are screened by algorithms. But it can also refer to practitioners who directly work with the AI system, such as doctors who use AI based image classifiers to detect skin cancer. As we will outline in more detail in the next section, users can exploit the shortcomings of AI systems for their own benefit.

# CORRUPT AI

The report proposes the following definition of corrupt AI:

*Abuse of AI systems by (entrusted) power holders for their private gain.*

When power holders, whether they are in public or private sectors (see Box 1 below) abuse AI systems to achieve personal benefits, this constitutes corruption. Due to the constitutive element of power, corrupt AI is not "just another form of crime" but a particularly challenging form of crime. Namely, in increasingly digital societies, those with the code and the data become more powerful.[28] AI can thus consolidate and even exacerbate existing power imbalances.[29] Those with the power to change corruption have often little incentive to do so because they benefit from it. Conversely, those with an incentive to change corruption – for example, the victims – have little power to do so. AI in the hands of the powerful only increases this trend. Autocratic regimes and a weak rule of law further exacerbate the risk that AI in these societies will be deployed in a corrupt manner, such as by a political or economic leadership seeking self-enrichment or a consolidation of their grip on power via the illegitimate suppression of opposition.[30] Over the past decades, political systems worldwide have been shifting towards more autocratic forms of government.[31] As such, it is reasonable to anticipate that, in many countries, AI is being developed and deployed with corrupt intentions in mind.

Why corrupt AI is particularly problematic becomes most apparent when examining a few illustrative examples. We categorise these examples according to whether:

i) the AI system is intentionally designed for corrupt purposes;
ii) the code or training data of existing AI systems are manipulated to achieve corrupt goals; or
iii) an AI system is applied in a corrupt fashion.

### Box 1 - Distinction between public and private corruption forms

The corruption literature distinguishes different forms of corrupt behaviour in the public or the private sector.

Public corruption refers to abuses of entrusted power for private gain in the public sector. Public officials are directly entrusted with the power to act in the public interest and are expected to do so impartially. Public corruption includes high-ranking heads of states embezzling public funds but also lower-ranking public officials like traffic police officers requesting bribes.

Private corruption refers to abuses of forms of power not entrusted within the public sector. Private sector companies are entrusted with power indirectly through the government issuing permits and licences. These licences and permits are issued based on certain standards and expectations of impartiality. For example, cheating your patients, users and clients is a clear breach of that expectation, especially when delivering basic services, such as medical services. Concrete examples are bribery of hiring managers or doctors and nurses.

## CORRUPT DESIGN OF AI

An example of a corrupt design of AI is when politicians and other power holders commission the generation of hyper-realistic deepfakes to discredit and intimidate (political) opponents to cement their power. With advances in the subfield of computer vision, it has become increasingly easy to produce such deepfakes.[32] Audio-visual content no longer serves as the gold standard for establishing veracity online. Empirical evidence indicates that people can no longer reliably detect such AI enabled synthetic media but remain overconfident in their abilities to do so.[33] Furthermore, political actors can strategically design AI systems for computational propaganda.[34] That is, social media bots can impersonate human users to push political agendas and manipulate public opinion on contentious societal issues like corruption.[35]

A major concern is that fake information generally travels faster and permeates social media networks deeper than accurate information.[36] AI systems that are intentionally designed to manipulate often put attackers at an advantage over defenders, notably on social media sites.[37] Not surprisingly then, in

recent elections, deepfakes have been used to sully the reputation of political candidates, especially female ones.[38] In such cases, politicians and other power holders commission the design of AI systems to manipulate broad audiences to remain in office, which is a form of private gain, especially in kleptocratic settings where public office is abused to plunder state coffers.

## CORRUPT MANIPULATION OF AI

Corrupt AI does not just occur when an AI system is designed with malicious intent. It can also take place when the vulnerabilities of existing – otherwise beneficial AI systems – are exploited. Consider, for example, a manipulated triage algorithm deciding on the distribution of ventilators during the COVID-19 pandemic. Many countries decided that a patient's survival probability should be the dominant factor when allocating scarce resources in emergencies. Even though AI systems are (to the best of our knowledge) currently not used to make such decisions, their potential use to allocate scarce medical resources such as ventilators is being openly and critically discussed.[39]

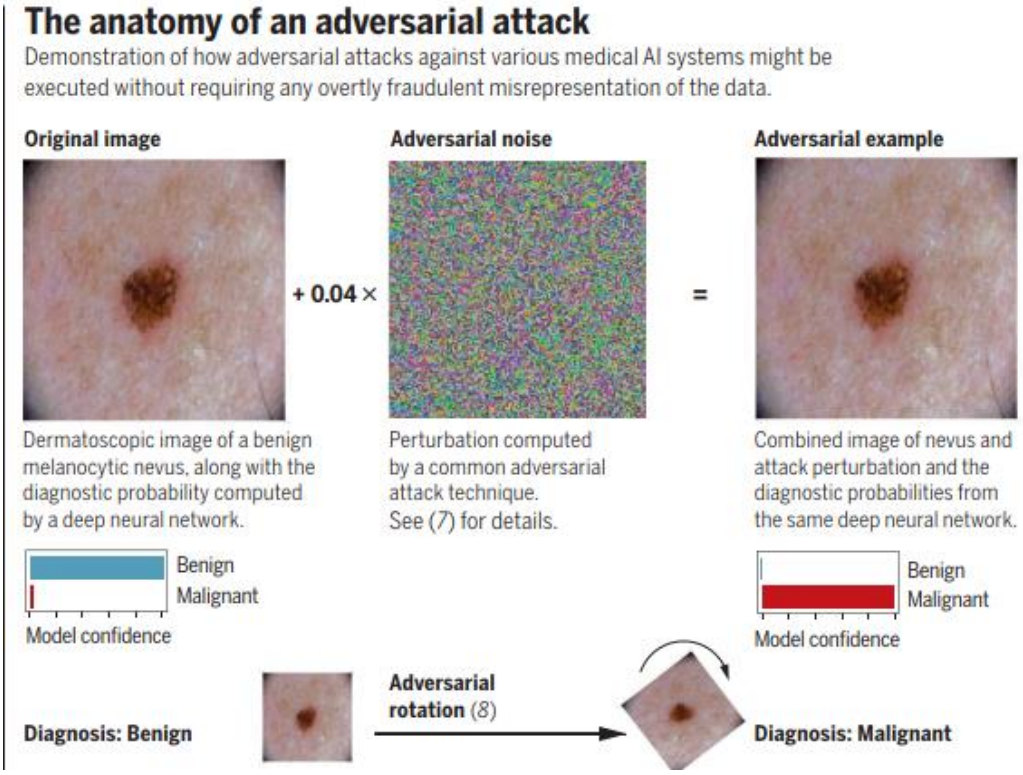
Dishonest data scientists tasked with designing an AI based system to predict the likelihood of a patient's survival could tweak the algorithms to favour themselves, their peer group or those who can afford treatment. They could intentionally bias the AI system so that it systematically discriminates based on demographics such as age or race. This way, the data scientists could ensure that they and their peers receive the best treatment in case they need emergency care. Such cases of manipulating AI models to systematically favour a specific group have been termed algorithmic capture.[40]

Another example of algorithmic capture stems from the domain of algorithmic hiring. Here, AI models are used to decide who to hire or to admit to universities. These decisions are based on automated analyses of applicants' resumes, test scores, interview footage, and so on. As Rahwan outlines:

> **"A programmer may design the AI to implement a kind of 'affirmative action' policy, increasing the representation of particular groups based on their gender, race, nationality, etc. Another programmer may design the AI to be fair in that it ignores such demographic factors. A third programmer may alter the algorithm to subtly favour 'native accents' in the interview video, excluding immigrants even if the job does not require such language skills.**
>
> Rahwan (2022)[41]

Other conceivable examples of algorithmic capture can entail corrupted e-procurement or fraud detection algorithms. These examples of algorithmic capture require a one-time manipulation of the AI system by power holders to reap benefits over a long period, possibly affecting millions of people. In this case, the corrupt manipulation of AI can be easily scaled to affect large groups of people. Other forms of corrupt manipulation of AI require the corruptor to tweak the algorithm on a case-by-case basis, for example, in the health sector. The same AI system that can improve the medical treatment of millions of patients can be manipulated for private gain by doctors. Many machine learning systems are vulnerable to so-called adversarial attacks.[42] The same image classification algorithms that can detect lung cancer based on x-ray images with high precision can be fooled (see illustration in Figure 2). Forcing the system to misclassify an image often requires only changing pixels or the angle at which the picture is taken. These perturbations to fool the image classification algorithm are usually imperceptible to the human eye. Malicious actors can therefore trick AI systems into seeing illnesses that do not, in fact, exist.[43] Those entrusted to operate such AI systems – for example, doctors and other hospital staff – can manipulate the AI to influence billing or insurance software to maximise their revenue.

**Figure 2.** Illustration of how image classification algorithms can be manipulated for private gain. Source: Finlayson et al. 2019.




Such adversarial attacks can occur by tweaking the training data or the algorithm and are not specific to image classification models but, in fact, have been demonstrated for most classes of machine learning algorithms.[44] Other forms of corruption can occur when these vulnerabilities are exploited. For example, in language models, substituting synonyms can be sufficient to fool such algorithms.[45] One example of the corrupt use of such adversarial attacks is when hiring officers sell the information about these vulnerabilities of a text classification model to applicants for bribes. While, at its core, it is not an entirely new form of corruption (bribery), the opacity of the algorithmic system involved renders the detection particularly difficult.

## CORRUPT APPLICATION OF AI

Some AI systems are created for a particular (benign) aim but are then repurposed for corruption. Consider the use of microtargeting – the use of advertising targeted to small groups of people based on their identified preferences. Such advertisements often occur on social media platforms. In recent elections, investigative journalists have documented cases where political power holders abused their public office and funds to run such microtargeted advertising campaigns to promote their parties.[46]

Other examples where power holders have abused AI based tools for private gain are the revelations around the Pegasus files and the NSA. Here, politicians and other power holders misappropriated existing AI systems to solidify or expand their power by surveilling, threatening and intimidating political or business rivals. A famous example stems from the large-scale revelations of the Pegasus project. The official use of the software is for counter-terrorism purposes. Yet, as is often the case with such large-scale digital surveillance technology, it can be abused for private gain. Evidence gathered by investigative journalists and civil society organisations implies that governments use mass-scale surveillance systems to spy on journalists, fellow politicians and other political opponents.[47] The Pegasus spyware grants unrestricted access to breached phones. It even

allows activating the microphone and recording private conversations.[48] Similar cases have been documented with the NSA overstepping its boundaries and using access to private phone calls for political and private advantages.[49] Such sophisticated spyware thus enables intrusion and data collection of unprecedented scope and depth.

What these examples highlight are some features that are common to all forms of corruption – the abuse of power. They also show that using AI to engage in corruption affects some of the dynamics of corruption itself. Classic forms of corruption are transformed when AI systems are used as a means to perpetrate corrupt acts.

## UNIQUE FEATURES OF CORRUPT AI

The next section distils the key technical and human factors that render corrupt AI unique and set it apart from other digital technologies. These technical factors are AI's ability to act autonomously, its opaque workings and personalisation towards the recipient on a large scale. These unique technical factors of AI, in turn, cause several human risk factors, namely, diffusion of responsibility, plausible deniability and psychological distancing of the victims.

### Technical factors

The first key feature lies in AI's ability to act with autonomy. Instead of executing tasks in a strictly predefined way, machine learning algorithms can act autonomously and unpredictably. The fact that these systems can, under specific circumstances, act autonomously is a crucial reason why many AI researchers propose to view such systems as "agents".[50] This autonomy imbues AI systems with greater responsibility, but it also undermines trust as people feel a lack of control over the outcomes.[51] A unique corruption risk emanates from this autonomy as AI systems can become corrupt actors, for example, when power holders programme bots to influence public opinions.[52]

AI systems often lead to opacity. Outputs by complex algorithms defy simple explanations, earning them the name of "black box" algorithms. Even to the programmers, the decisions reached are not easily traceable and, at times, entirely incomprehensible.[53] Also, the datasets on which the algorithm is trained often remain hidden from public scrutiny. As the algorithmic capture examples illustrate, malicious actors could manipulate code and/or training data for private gain. Detecting such cases is notoriously difficult.

Another key feature of AI that is important for understanding its potential misuse is the unique combination of personalisation and scalability. Scaling up through AI often comes at low marginal costs and allows content to be personalised to a given recipient. As we outlined in the deepfake example, corrupt actors can thereby reach unprecedented audiences at a rapid speed. Microtargeting further allows tailored content to be delivered to people. At times, the same political parties even use contradicting messages for different audience segments.[54]

When used maliciously to deceive, AI technology, thus, has a powerful manipulative force with unparalleled reach. In other words, corrupting a human often involves only a few transactions, while corrupting an instance of AI technologically can systematically distort millions of transactions. These features turn AI systems into attractive technologies for bad actors to exploit. From a classic economic cost-benefit perspective, they offer higher rewards by providing effective manipulative tools while simultaneously reducing the risk of detection (and thus punishment) through their opaque workings and anonymity.

### Human factors

Besides technical factors, AI also bears human risk factors for corruption. Firstly, corruption through AI systems often increases the diffusion of responsibility. Namely, the constraints on corruption are lowered when AI is involved in committing corrupt acts because diffusion of responsibility makes the detection and sanctioning of this corrupt behaviour less likely. Diffusion of responsibility is a classic phenomenon in behavioural research.[55] In cases of misconduct, people seek to deflect blame towards other co-culprits.[56] These co-culprits have classically been fellow humans. For example, in bribery transactions, it is common for those involved to accuse their partner of having instigated the transaction.[57] By now, people do not just deflect blame on fellow humans but also on AI systems.[58] In some ways, this is even more appealing than blaming a fellow human.[59] An AI system cannot (yet) contradict the accusation. For example, the doctor who manipulates the AI image classifier to boost profit might deflect the blame to the seemingly faulty algorithm.

Also, establishing clear culpability is particularly difficult when AI is used to engage in corruption. It is often infeasible or even impossible to detect whether someone tinkered with the data or algorithms, which increases plausible deniability. In many instances – particularly those we are not aware of yet – the manipulative involvement of humans in the AI decision-making process remains hard to detect, enabling the corrupt actors to deny their culpability. As a clear breadcrumb trail to the manipulation of the AI system is often missing, it is easy for bad actors to deny their involvement. To this day, it is unclear which government (officials) used the data obtained via the Pegasus software. For policy-makers, the involvement of AI therefore raises new challenges for establishing liability and prosecution.[60]

Using AI to engage in corruption also increases the psychological distance from the victims. Victims of corruption are already often vague and distant; public discourse sometimes labels some forms of corruption a victimless crime.[61] However, it is undeniable that people suffer from corrupt acts in the end; for instance, earthquake casualties who would have survived if the building safety inspectors were not bribed to look the other way.[62] Having AI systems as a tool to engage in corruption arguably increases the psychological distance to victims even further.

# RECOMMENDATIONS

Those regulating new technology often face the Collingridge dilemma: when new technologies arise on the scene, it is typically accompanied by two competing concerns. "On one hand, regulations are difficult to develop at an early technological stage because their consequences are difficult to predict. On the other hand, if regulations are postponed until the technology is widely used, then the recommendations come too late".[63]

While some forms of corrupt AI are already documented (for example, adversarial attacks on the health system) or have even led to policy responses at the EU level (for example, Pegasus project), evidence for other cases of corrupt AI (for example, involving auditors) is scarce at the moment.

That does not mean corrupt uses of AI cannot become a threat to be reckoned with in the near future. In fact, we might simply not know about many other already existing cases because of the outlined difficulty of identifying them. Currently, not much (policy) attention is placed on corrupt AI. The following recommendations on how to meet the emerging threat of corrupt AI are categorised according to whether they address regulatory, technical or human factors.

## REGULATORY FACTORS

In recent years, lawmakers and international organisations have called for new regulations for AI. It is contested whether AI systems force legal codebooks to be updated or entirely overhauled.[64] Establishing guidelines for the ethical development and implementation of AI plays an integral role in most regulatory approaches. For instance, the High-Level Expert Group on AI by the European Commission, the OECD and UNESCO put forth regulatory guidelines outlining different principles of ethical AI, such as transparency, fairness and accountability.[65] Comparative research reveals much overlap in the suggested principles of ethical AI.[66] Hagendorff finds that the requirements for accountability, privacy and fairness can be found in 80 per cent of the 22 guidelines he analysed.[67] A systematic review by Jobin and colleagues identifies five key principles (transparency, justice and fairness, non-maleficence, responsibility and privacy) referred to in more than half of the guidelines.[68]

While these ethical guidelines are crucial to establishing sound regulatory frameworks for emerging AI technologies, for the most part, they have neither been translated into binding legislation nor have they specifically highlighted the danger of corrupt AI. More recent legislative acts proposed by the European Commission suggest that legislators become more aware of the potential misuse of AI. For instance, the Commission's 2021 Artificial Intelligence Act warns, "Aside from the many beneficial uses of artificial intelligence, that technology can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices".[69] Robust legislation is needed to alleviate the misuse of AI technologies and the potentially harmful societal consequences associated with implementing AI.

For instance, due to its autonomous abilities, AI raises new challenges for establishing moral and legal culpability. The companies behind algorithms often promote their products to solve a particular problem – in one documented instance, to detect cheaters in online exams – but shift final responsibility back to the human principal who purchases the product – in the above case, the college.[70] Establishing legal frameworks that define liability for AI systems is a general challenge and also one for preventing AI corruption. Namely, when legal responsibility remains unsolved, corruption risks are heightened as people can hide behind the code. Open data and code via creative commons licences facilitate algorithmic auditing and protect those who blow the whistle about AI corruption.

## TECHNICAL FACTORS

### Data and code transparency

Being able to conduct code audits requires data and code to be openly available.[71] Private companies in particular are often reluctant to publicise their data and code for proprietary reasons. Therefore, there

is a lack of infrastructure to keep them in check and detect shortcomings in their data or models. Hence, data and code transparency presents an important foundational step to enable accountability and avoid AI corruption. Algorithmic transparency describes the principle of making the factors that influence the decision of an AI system transparent to the relevant stakeholders.[72] Similarly, algorithmic accountability stipulates that those people and institutions employing AI systems must be accountable for the consequences.[73] Whether transparency around AI system data and code actually leads to algorithmic accountability depends on what and how information is made available.[74] For a more in-depth discussion on which aspects to consider for algorithmic transparency, see Kossow (2021).[75]

### Facilitating model audits

Another promising approach focuses on implementing rigorous and independent audits.[76] Independent audits (for example, by civil society organisations like the Algorithmic Justice League or Algorithm Watch) can ensure that algorithms are designed to adhere to the ethical principles outlined in most existing regulatory guidelines. They can establish safeguards against the intended misuse of AI for corrupt and other illegal/immoral activities, as well as unintended consequences associated with implementing AI in societal contexts. By opening the "black box", audits can provide the much needed transparency for scrutinising the use of AI in the private and public sectors.[77]

Audits require not only transparent data and code but also two other technical features. The first deals with a current lack of emphasis on the importance of continuous quality checks. Machine learning operations (MLOps) – a set of practices to deploy and maintain machine learning models reliably and efficiently in production – tend to be neglected.[78] There is currently a gold rush in developing new machine learning models and releasing them to the market, yet only a few companies invest in their maintenance and constant quality checks.

The second technological feature to reduce the risk of AI corruption lies in making such code audits easier. One concrete way is to facilitate the interoperability of machine learning programming. Currently, data scientists use different programming languages for machine learning models, most commonly PyTorch or TensorFlow. Auditing the code is more tedious across different programming languages. Efforts such as the ONNX open format help to improve interoperability and make code audits easier. It does so by defining a "common set of operators – the building blocks of machine learning and deep learning models – and a common file format to enable AI developers to use models with a variety of frameworks, tools, runtimes, and compilers".[79] More emphasis on MLOps and interoperable coding languages could also help to detect biases in existing machine learning models – whether unintentionally or intentionally corrupt ones.

Moreover, testing and developing MLOps algorithms that are resilient to adversarial attacks pose an important challenge. For example, training algorithms with exposure to adversarial examples can help to reduce the vulnerabilities of adversarial attacks.[80] Also, where it is possible to manipulate or fool AI models, storing so-called fingerprints (digital imprints of users) as so-called hashes can facilitate audits. Namely, such digital stamps help create a breadcrumb trail to the manipulative action. Technical guardrails like these can reduce the risks of AI abuse by power holders for private gains.

## HUMAN FACTORS

### Ethics training

Data scientists and code auditors have become important stakeholders in the implementation of AI systems. This ascent to a position of great power has happened very rapidly, in both the public and the private sectors. In contrast to classic professions in power, such as politicians, police officers or doctors, there are no professional codes of conduct, let alone specific ones for anti-corruption, in place. At the same time, ethics training for programmers and data scientists is among the most common recommendations to ensure ethical and responsible AI.[81] Much like classic approaches to fighting corruption, this idea rests on the assumption that raising decision-makers' awareness and sensitivity to the ethical repercussions of their actions helps avoid harmful outcomes. Currently, such pieces of training are generally uncommon. A recent report on the issue suggests that "when asked about the topics being taught to data science/ML students, only 17% and 22% of educators responded that they were teaching about ethics or bias, respectively".[82] Hence, a starting point would be to sensitise data scientists, programmers and code auditors to AI corruption risks. This can happen through professional training, codes of conduct or compliance guidelines. Such efforts should be

accompanied by a rigorous evaluation of their effectiveness.

## Overcoming reduced whistleblower capacity

In many instances, AI systems are used to replace humans. As companies employ increasingly fewer humans to oversee crucial tasks, the whistleblower capacity within the institutions decreases. Whistleblowing requires independent actors to speak out against their employees. AI algorithms are much more likely to have incentives aligned with the company/institution that implements them. Therefore, AI systems – in their current form – have no internal reporting or whistleblowing capacities.

Two factors are at play here. First, replacing humans with AI reduces the absolute number of those who can engage in reporting. Second, introducing AI systems might also reduce the willingness and confidence of those left to blow the whistle. People might have (overly) positive views about the performance of the AI systems, not suspecting that they can go rogue.[83] The fact that the algorithmic processes of AI systems are often opaque further reduces whistleblowing capacities. Hence, raising awareness about these two reductions in reporting and whistleblowing capacities marks an important step for sustaining the possibility of people speaking up against (AI) corruption.

# CONCLUSION

The impact of AI on societies around the globe continues to grow, while the rising capabilities of AI shift existing power structures. In a digital age, power resides with those who have the code and the algorithms – currently mostly large tech companies and governments. Corrupt AI occurs when power holders abuse this power for their private gain. This paper highlights that they can do so either by designing, manipulating or applying AI systems. Making AI systems more resilient against corruption risks requires novel safeguards.

Here, we call on policy-makers, programmers, private companies and civil society organisations to address three main aspects:

(1) Develop innovative regulatory frameworks that support the ethical design and implementation of AI as well as mandating model audits.
(2) Facilitate such audits by ensuring transparent code and data as well as the interoperability of different programming languages.
(3) Sensitise new powerful actors like data scientists and programmers to AI ethics and anti-corruption through training and codes of conduct.

# ENDNOTES


[1] Acemoglu, D. 2021. Harms of AI. NBER Working Paper, 29247 Sept 2021. https://doi.org/10.3386/w29247.

[2] Brynjolfsson, E., Rock, D., & Syverson, C. 2017. Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. No. 24001. National Bureau of Economic Research. https://doi.org/10.3386/w24001

[3] Gomes, C., Dietterich, T., Barrett, C., Conrad, J., Dilkina, B., Ermon, S., Fang, F., Farnsworth, A., Fern, A., Fern, X., Fink, D., Fisher, D., Flecker, A., Freund, D., Fuller, A., Gregoire, J., Hopcroft, J., Kelling, S., Kolter, Z., … Zeeman, M. L. 2019. Computational sustainability: computing for a better world and a sustainable future. Communications of the ACM, 62(9), 56–65.

[4] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.

[5] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., … & Greene, C. S. 2018. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface, 15(141), 20170387.

[6] Hutson, M. 2021. Robo-writers: The rise and risks of language-generating AI. Nature, 591(7848), 22–25.

Köbis, N. C., & Mossink, L. D. 2021. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior, 114(106553), 106553.

[7] Aarvik, P. 2019. Artificial Intelligence a promising anticorruption tool in development settings. No. 2019:1. U4 Anti-Corruption Resource Center.

Adam, I., & Fazekas, M. 2021. Are emerging technologies helping win the fight against corruption? A review of the state of evidence. Information Economics and Policy, 100950.

Köbis, N. C., Starke, C., & Rahwan, I. 2022. The promise and perils of using artificial intelligence to fight corruption. Nature Machine Intelligence. https://doi.org/10.1038/s42256-022-00489-1

[8] López-Iturriaga, F. J., & Sanz, I. P. 2018. Predicting public corruption with neural networks: An analysis of Spanish provinces. Social Indicators Research, 140(3), 975–998.

[9] De Blasio, G., D'ignazio, A., Letta, M., Balassone, F., Cerqua, A., Cerulli, G., Charpentier, A., Gianluca, M., Esposito, E., Flachaire, S., Mocetti, P., Montalbano, A., & Muscarnera, L. 2020. Predicting corruption crimes with machine learning. A study for the Italian municipalities. http://www.diss.uniroma1.it/sites/default/files/allegati/DiSSE_deBlasioetal_wp16_2020.pdf

[10] Lavigne, S., Clifton, B., & Tseng, F. 2017. Predicting financial crime: Augmenting the predictive policing arsenal. Preprint at https://arxiv.org/abs/1704.07826

[11] Forjan, J., Köbis, N. C., & Starke, C. 2022. Using artificial intelligence to fight corruption: Expert interviews on the potentials and limitations of existing approaches. In A. Mattoni (Ed.), Digital Media and Anticorruption. Routledge.

Odilla, F. 2021. Bots against corruption: Exploring benefits and limitations of AI-based anti-corruption technology. International Seminar Artificial Intelligence: Democracy and Social Impacts. https://www.academia.edu/download/67395812/FO_AI_Bots_Against_Corruption_2May2021.pdf

[12] Breslow, S., Hagstroem, M., Mikkelsen, D., & Robu, K. 2017. The new frontier in anti–money laundering. https://www.mckinsey.de/~/media/McKinsey/Business%20Functions/Risk/Our%20Insights/The%20new%20frontier%20in%20anti%20money%20laundering/The-new-frontier-in-anti-money-laundering.pdf

[13] Christian, B. 2020. The Alignment Problem: Machine Learning and Human Values. W. W. Norton & Company.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. No date. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. Big Data and Society. https://arxiv.org/abs/2103.12016

[14] Buolamwini, J., & Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Vol. 81, pp. 77–91). PMLR.

[15] King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. 2020. Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. Science and Engineering Ethics, 26(1), 89–120.

Köbis, N. C., Bonnefon, J.-F., & Rahwan, I. 2021. Bad machines corrupt good morals. Nature Human Behaviour, 5(6), 679–685.

[16] Damiani, J. 2019. A voice deepfake was used to scam a CEO out of $243,000. Forbes Magazine. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/

[17] Calvano, E., Calzolari, G., Denicolò, V., Harrington, J. E., Jr, & Pastorello, S. 2020. Protecting consumers from collusive prices due to AI. Science, 370(6520), 1040–1042.

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. 2020. Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. Science and Engineering Ethics, 26(1), 89–120.

[18] Ferrara, Emilio. 2015. "'Manipulation and Abuse on Social Media' by Emilio Ferrara with Ching-Man Au Yeung as Coordinator." SIGWEB Newsl., 4, , no. Spring (April): 1–9.

[19] Kipnis, D. 1976. The powerholders. Chicago: University of Chicago Press, p. 83

[20] Nichols, P. M. 2019. Bribing the machine: Protecting the integrity of algorithms as the revolution begins. American Business Law Journal, 56(4), 771–814.

[21] Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Boujemaa, N., Bauer, W., Bouarfa, L., Chatila, R., Coeckelbergh, M., Dignum, V., Gagné, J.-F., Goodey, J., Haddadin, S., Hasselbalch, G., Heintz, F., Hidvegi, F., Höckner, K., Jégo-Laveissière, M.-N., … Van Wynsberghe, A. 2019. Building trust in human-centric AI. European Commission.

[22] Mitchell, M. 2019. Artificial Intelligence: A Guide for Thinking Humans. Penguin UK.

[23] Mitchell, M. 2019. Artificial Intelligence: A Guide for Thinking Humans. Penguin UK.

[24] Misuraca, G., van Noordt, C., & Boukli, A. 2020. The use of AI in public services: Results from a preliminary mapping across the EU. Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, 90–99.

Starke, C., & Lünich, M. 2020. Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. Data & Policy. https://doi.org/10.1017/dap.2020.19

Wirtz, B. W., Weyerer, J. C., & Geyer, C. 2019. Artificial intelligence and the public sector: Applications and challenges. International Journal of Public Administration, 42(7), 596–615.

[25] Szymielewicz, K. 2020. Black-boxed politics: Katarzyna szymielewicz. Medium. https://medium.com/@szymielewicz/black-boxed-politics-cebc0d5a54ad

[26] Barocas, S., & Selbst, A. D. 2016. Big data's disparate impact. https://doi.org/10.2139/ssrn.2477899

[27] Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. 2018. Why we need to audit algorithms. Harvard Business Review. https://hbr.org/2018/11/why-we-need-to-audit-algorithms

[28] Kalluri, P. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature, 583(7815), 169.

[29] Köbis, N. C., Starke, C., & Rahwan, I. 2022. The promise and perils of using artificial intelligence to fight corruption. Nature Machine Intelligence. https://doi.org/10.1038/s42256-022-00489-1

[30] Köbis, N. C., Starke, C., & Rahwan, I. 2022. The promise and perils of using artificial intelligence to fight corruption. Nature Machine Intelligence. https://doi.org/10.1038/s42256-022-00489-1

[31] Alizada, Cole, Gastaldi, Grahn, & Hellmeier. 2021. Autocratization turns viral. Democracy report 2021. V-Dem Institute.

[32] Groh, M., Epstein, Z., Firestone, C., & Picard, R. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences, 119(1), e2110013119

[33] Groh, M., Epstein, Z., Firestone, C., & Picard, R. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences, 119(1), e2110013119

Köbis, N.C., Doležalová, B. et al. 2021. Fooled twice: People cannot detect deepfakes but think they can. iScience, vol.24(11)

[34] DiResta, R. 2018. Computational propaganda: If you make it trend, you make it true. The Yale Review, 106(4), 12–29.

[35] Salge, C. A. D. L., & Karahanna, E. 2018. Protesting corruption on twitter: Is it a bot or is it a person? Academy of Management Discoveries, 4(1), 32–49.

[36] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

[37] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. 2016. The rise of social bots. Communications of the ACM, 59(7), 96–104. https://doi.org/10.1145/2818717

[38] Diakopoulos, N., & Johnson, D. 2021. Anticipating and addressing the ethical implications of deepfakes in the context of elections. New Media & Society, 23(7), 2072–2098.

[39] Helbing, D., Beschorner, T., Frey, B., Diekmann, A., Hagendorff, T., Seele, P., Spiekermann-Hoff, S., van den Hoven, J., & Zwitter, A. 2021. Triage 4.0: On death algorithms and technological selection. is today's data-driven medical system still compatible with the constitution? Journal of European CME, 10(1), 1989243.

Lünich, M., & Kieslich, K. 2022. Exploring the roles of trust and social group preference on the legitimacy of algorithmic decision-making vs. human decision-making for allocating COVID-19 vaccinations. AI & Society, 1–19.

[40] Rahwan, I. 2022. Prevent algorithm capture. Evil AI Cartoons. https://www.evilaicartoons.com/archive/prevent-algorithm-capture

[41] Rahwan, I. 2022. Prevent algorithm capture. Evil AI Cartoons. https://www.evilaicartoons.com/archive/prevent-algorithm-capture

[42] Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. 2019. Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289.

[43] Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. 2019. Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289.

Metz, & Smith. 2019. Warnings of a dark side to AI in health care. The New York Times. https://www.nytimes.com/2019/03/21/science/health-medicine-artificial-intelligence.html

[44] Biggio, B., & Roli, F. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.

[45] Li, J., Ji, S., Du, T., Li, B., & Wang, T. 2018. TextBugger: Generating adversarial text against real-world applications. In arXiv [cs.CR]. arXiv. http://arxiv.org/abs/1812.05271

[46] TargetLeaks. 2019. TargetLeaks. https://targetleaks.de/

[47] Pegg, D., & Cutler, S. 2021. What is Pegasus spyware and how does it hack phones? The Guardian. http://www.theguardian.com/news/2021/jul/18/what-is-pegasus-spyware-and-how-does-it-hack-phones

[48] Kirchgaessner. 2021. New evidence suggests spyware used to surveil Emirati activist Alaa Al-Siddiq. The Guardian. https://amp.theguardian.com/world/2021/sep/24/new-evidence-suggests-spyware-used-to-surveil-emirati-activist-alaa-al-siddiq

[49] Greenwald, G. 2014. No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State. Henry Holt and Company.

[50] Floridi, L. 2016. Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 374(2083). https://doi.org/10.1098/rsta.2016.0112

[51] Alaieri, F., & Vellino, A. 2016. Ethical Decision Making in Robots: Autonomy, Trust and Responsibility. Social Robotics, 159–168.

[52] Salge, C. A. D. L., & Karahanna, E. 2018. Protesting corruption on twitter: Is it a bot or is it a person? Academy of Management Discoveries, 4(1), 32–49.

[53] Stoyanovich, J., Van Bavel, J. J., & West, T. V. 2020. The imperative of interpretable machines. Nature Machine Intelligence, 2(4), 197–199.

[54] TargetLeaks. 2019. TargetLeaks. https://targetleaks.de/

[55] Darley, J. M., & Latané, B. 1968. Bystander intervention in emergencies: diffusion of responsibility. Journal of Personality and Social Psychology, 8(4), 377–383.

[56] Bartling, B., & Fischbacher, U. 2011. Shifting the blame: On delegation and responsibility. The Review of Economic Studies, 79(1), 67–87.

[57] Köbis, N. C., van Prooijen, J. W., & Righetti, F. 2016. Prospection in individual and interpersonal corruption dilemmas. Review of General Psychology, 20(1), 71–85.

[58] Hohenstein, J., & Jung, M. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. Computers in Human Behavior, 106, 106190.

[59] Köbis, N. C., & Mossink, L. D. 2021. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior, 114(106553), 106553.

[60] Citron, D. K. 2007. Technological due process. Wash. UL Rev. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/walq85&section=38&casa_token=gJE2h5WctA4AAAAA:GQMeZEJmmMy-dktgSxjVUsnA08FNvMzQq7XlDT5SFBqJwjLOOg5XsXpEyTWQtGD2peIeOAYHeg

[61] Köbis, N. C., van Prooijen, J. W., & Righetti, F. 2016. Prospection in individual and interpersonal corruption dilemmas. Review of General Psychology, 20(1), 71–85.

[62] Ambraseys, N., & Bilham, R. 2011. Corruption kills. Nature, 469(7329), 153–155.

[63] Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., & Rahwan, I. 2020. Crowdsourcing moral machines. Communications of the ACM, 63(3), 48–55, p. 53

[64] Gutierrez, C. I. 2020. The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the US. https://papers.ssrn.com/abstract=3649707

[65] European Commission. 2021. Proposal for a regulation of the EUROPEAN PARLIAMENT and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM 2021 206 final). EU Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

OECD. No date. Recommendation of the council on artificial intelligence. OECD. Retrieved July 8, 2022, from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

UNESCO. 2021. Draft text of the Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000377897

[66] Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120.

Jobin, A., Ienca, M., & Vayena, E. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

[67] Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120.

[68] Jobin, A., Ienca, M., & Vayena, E. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

[69] European Commission. 2021. Proposal for a regulation of the EUROPEAN PARLIAMENT and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM 2021 206 final). EU Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, p. 21

[70] Hill, K. 2022. Accused of cheating by an algorithm, and a professor she had never met. The New York Times. https://www.nytimes.com/2022/05/27/technology/college-students-cheating-software-honorlock.html

[71] Kossow, N., Windwehr, S., & Jenkins, M. 2021. Algorithmic transparency and accountability. Transparency International. https://www.jstor.org/stable/pdf/resrep30838.pdf

[72] Diakopoulos, N., & Koliska, M. 2017. Algorithmic transparency in the news media. Digital Journalism, 5(7), 809–828.

[73] Diako N. Diakopoulos. Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. Digital Journalism. 3 (3), 2015.

[74] Kossow, N., Windwehr, S., & Jenkins, M. 2021. Algorithmic transparency and accountability. Transparency International. https://www.jstor.org/stable/pdf/resrep30838.pdf

[75] Kossow, N., Windwehr, S., & Jenkins, M. 2021. Algorithmic transparency and accountability. Transparency International. https://www.jstor.org/stable/pdf/resrep30838.pdf

[76] Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. 2018. Why we need to audit algorithms. Harvard Business Review. https://hbr.org/2018/11/why-we-need-to-audit-algorithms

[77] UK Government. 2022. Auditing algorithms: The existing landscape, role of regulators and future outlook. UK Government. https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook

[78] Kreuzberger, D., Kühl, N., & Hirschl, S. 2022. Machine learning operations (MLOps): overview, definition, and architecture. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/2205.02302

[79] ONNX. No date. Retrieved 29 May 2022, from https://onnx.ai/

[80] Biggio, B., & Roli, F. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.

[81] Weinstein, J., Reich, R., & Sahami, M. 2021. System error: Where big tech went wrong and how we can reboot. Hachette UK.

[82] Quiroga, H. 2017. Anaconda

[83] Scheer, K., Rabl, T., & Köbis, N. C. 2022. AI-based software as anti-corruption agent and employees' likelihood of whistleblowing. Proceedings: A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium, 2022(1), 15895.